ELSEVIER

Information Sciences 146 (2002) 67–73

INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL

www.elsevier.com/locate/ins

Oligonucleotide microarray data distribution and normalization

I.A. Sidorov [a,*], D.A. Hosack [b], D. Gee [a], J. Yang [b], M.C. Cam [c], R.A. Lempicki [b], D.S. Dimitrov [a]

[a] *National Cancer Institute, NIH, NCI-Frederick, MD 21702, USA*
[b] *SAIC, NCI-Frederick, MD 21702, USA*
[c] *Diabetes Branch, NIDDK, NIH, Bethesda, MD 20892-0842, USA*

Received 1 June 2001; received in revised form 1 November 2001; accepted 1 February 2002

Abstract

Variations in oligonucleotide microarray probe signals that result from various factors, including differences in sample concentrations, can lead to major problems in the interpretation of data obtained from different experiments. Normalization of such signals is typically performed by procedures involving division by a constant approximately determined by average signal intensities as, e.g., in the Affymetrix software. Here we show that Affymetrix oligonucleotide probe signal distributions can be fitted by using a superposition of two normal or two extreme distributions, and that by using such distributions we can normalize data with high accuracy (parametric algorithm). We also developed a second algorithm (nonparametric) based on ranking of signal intensities which gave equal or better normalization than the parametric one. These approaches have been used for normalization of three sets of data obtained from cancer cell lines, peripheral blood mononuclear cells from patients with HIV infections, and adipose cells from patients with diabetes, and others. Both, parametric and nonparametric normalization procedures, were found to be superior when compared to the standard global normalization approach [Affymetrix Microarray Suite User Guide. Version 4.0 (2000)]. These results suggest that the new approaches may be helpful for microarray data normalization especially for comparison of clinical data where inter-patient differences can be large and difficult to avoid.

[*] Corresponding author.

0020-0255/02/$ - see front matter © 2002 Published by Elsevier Science Inc.
PII: S0020-0255(02)00215-3

1. Introduction

Studies of large-scale gene expression based on oligonucleotide or cDNA microarray technology have become an important part of biomedical research over the last few years [2,3]. Gene expression in a high-density oligonucleotide microarray is measured by the signal intensities of probe pairs: perfect match, PM, and mismatch, MM. The number of these pairs of probes for each particular gene can vary significantly (from several to tens of probes). The average difference PM–MM or the logarithm of the ratio PM/MM is usually used to calculate gene expression.

Variations in oligonucleotide microarray probe signals, which arise from a variety of sources including differences in sample concentrations, can lead to major problems in the interpretation of data obtained from different experiments. To compare probe intensities for two or more arrays it is necessary to normalize them. Several different approaches have been used for normalization. The simplest method is the use of a single normalization factor for all probe signal intensities [1]. This factor can be calculated as a ratio of average differences PM–MM or log(PM/MM) for the two arrays under comparison. As has been previously noted [4], linear relation between intensities of different arrays does not hold in general and the distribution of low-intensity signals behaves differently from the distribution of high-intensity signals. To account for this discrepancy, a change point detection technique was applied wherein the entire set of intensities was divided into two blocks so that linear regression could be used effectively. Another approach is based on a given set of "housekeeping" genes (genes believed to be equally expressed for two different experiments). Two criteria used to measure the quality of normalization were proposed [5]: (1) a minimum of PM–MM difference variance across a series of arrays and (2) the stability of expression ratios in simulated data.

Here we show that Affymetrix oligonucleotide probe signal distributions can be fitted by using a superposition of two normal or two extreme distributions, and that by using such distribution we can normalize data with high accuracy (parametric algorithm). We also developed a second algorithm (nonparametric) based on ranking of signal intensities which gave equal or better normalization than the parametric one. Differences between values of normalized and model sample histograms as well as average correlation coefficients between probe intensities before and after normalization for all genes were used as criteria for the quality of normalization.

2. Probe signal intensity distributions

The probe signal intensities can vary significantly (several orders of magnitude). The distribution of the logarithm of probe signal intensities for one

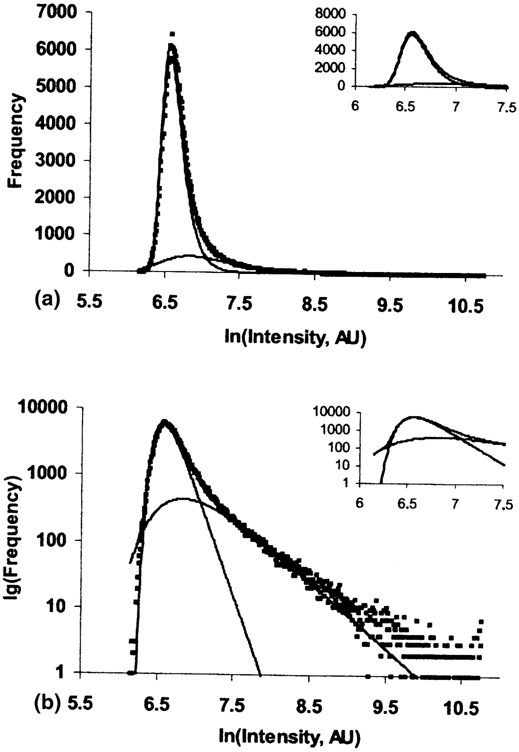

Fig. 1. Distribution of logarithm of probe intensitites in regular (a) and logarithm (b) scale.

sample is represented in Fig. 1. One can see that the distribution is nonsymmetrical and at higher intensities, the histogram exhibits a two-exponential behavior. The histogram data can be fitted by the nonsymmetrical probability distribution function (pdf) or by the combination of symmetrical and/or nonsymmetrical pdfs. To fit the data a combination of normal, gamma, largest extreme and Weibull pdfs [6] were tested. The best fitting was obtained when weighted sum of two log-extreme largest value pdfs was used (Fig. 1):

$$E(x,a,b) = \sum_{i=1}^{2} \frac{A_i}{b_i} \exp\left(\frac{a_i - \ln x}{b_i}\right) \exp\left(- \exp\left(\frac{a_i - \ln x}{b_i}\right)\right),$$

where a_i, b_i are the parameters of ith largest extreme value distribution and A_i are weighting factors ($i = 1, 2$). Introduction of a third term in the distribution function did not improve fitting significantly. Using these results, one can assume that the whole set of logarithm of signal intensities in a microarray can be read as a product of two random variables having the log-extreme largest value distribution with different parameters in general. The ith subset of variables consists of $N_i = NA_i/(A_1 + A_2)$ values, where N is the total number of probes in microarray.

3. Parametric normalization

Let us consider two sets of probe signal intensity values: array to be normalized and model array (array against which normalization will be done) and assume that the whole set of logarithm of signal intensities for each microarray can be generated by realization two log-extreme largest random variables with parameters a_i, b_i, $i = 1, 2$. Let us then transform the values within each subset of array to be normalized with respect to corresponding subsets of model array. The transformation for each subset is linear and can be calculated as

$$I_i^N \to b_i^M \frac{\ln I_i^N - a_i^N}{b_i^N} + a_i^M.$$

Here $i = 1, 2$ means the first and the second subset, respectively, and subscripts M and N denote model and array to be normalized, respectively. After this linear transformation, we will have two subsets of intensities distributed as largest extremes with the parameters a_i^M, b_i^M.

During the data transformation we have to know for each probe, whether it belongs to the first or to the second subset. This problem can be solved by introducing a decision function. Let us consider two values

$$q_i = p_i/(p_1 + p_2), \quad i = 1, 2,$$

where $p_i(x) = A_i \min\ (E_C(x, a_i, b_i),\ 1 - E_C(x, a_i, b_i))$ and $E_C(x, a, b)$, the cumulative distribution function of log-extreme largest value. Let us define q_i as the probability that x belongs to the ith subset of probes. One possible formula for the decision function is $[2 - q_1(x)] = i$, where: [*] is the nearest integer number and i is equal to 1 or 2 when x belongs to the to the first or second subset of probes, respectively. The other decision can be generated by the following rule: x belongs to the first subset if $r < q_1(x)$ and to the second one if $r \geqslant q_1(x)$, where $r =$ random variable uniformly distributed in the interval [0, 1].

4. Nonparametric normalization

Let us order the probe signal intensity values for both model and arrays to be normalized and consider the following procedure for aligning the two histograms. The frequency M_k for the kth interval of histogram $[X_{\min} + (k-1)X_D, X_{\min} + kX_D]$ depends on the three user-defined histogram parameters: minimum ($X_{\min}$) and maximum ($X_{\max}$) values and numbers of intervals N ($X_D = (X_{\max} - X_{\min})/N$, $k = 1, \ldots, N$). Let us transform M_k values of an array to be normalized with indexes from n_k to $n_k + M_k - 1(n_0 = 1)$ using the formula $\ln I_i^N \to (\ln I_i^N - b)/a$, where a, and b are the parameters transforming linearly the interval $[\ln I_{n_k}^M, \ln I_{n_k+M_k-1}^M]$ to the interval $[\ln I_{n_k}^N, \ln I_{n_k+M_k-1}^N]$ and repeat the calculation for all values of k $(n_{k+1} = n_k + M_k)$. When $k = N$, the resulting data set is normalized.

5. Quality of normalization algorithms

Two parameters were used to estimate the quality of the normalization methods. The first one is the sum of square of differences between the model and normalized histogram values (Q). For two identical sets of values, Q equals 0, and does not depend on the parameters of the histogram. Data transformation should not change significantly the proportion between different probe signal intensities for each particular gene before and after normalization. To measure this difference we used the second parameter for the quality of normalization: Pearson correlation coefficient (R) between signal intensity values before and after normalization for each gene. Average values and standard deviations of R for all genes as well as minimum and maximum values of R and number of genes having $R < 0.5$ were calculated for each microarray.

6. Results

By using the parametric and nonparametric procedures we normalized microarray data obtained from cancer cell lines, adipose cells from patients with diabetes, and peripheral blood mononuclear cells from patients with HIV infections (Fig. 2, 12 microarrays). Both, parametric and nonparametric normalization procedures, were better compared to the standard global normalization approach. One can see that for both approaches, the final distribution is very close to the distribution of model data set. The average values of R for genes before and after parametric and nonparametric normalization were close to 1.0 with typical standard deviation of about .01–.05, These results suggest that the new approaches could be helpful for microarray data normalization especially for comparison of clinical data where the interpatient differences can be large and difficult to avoid.

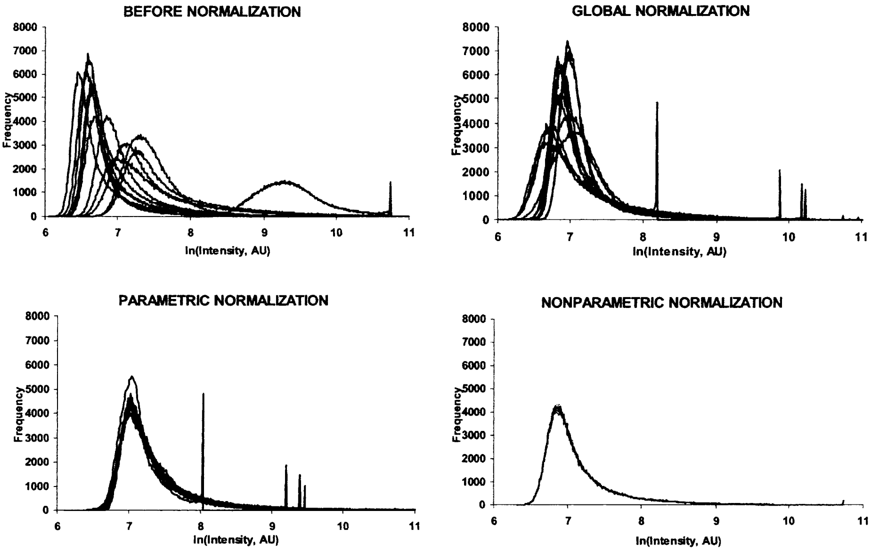

Fig. 2. Distribution of logarithm of probe intensities for 12 different experiments before and after global, parametric, and nonparametric normalization.

7. Discussion

We have developed two different algorithms for microarray data normalization which have been applied to clinical data. Both algorithms are based on the assumption that up- and down-regulated genes do not change significantly the histogram of probe signal intensity distributions.

The first algorithm (parametric normalization) requires data fitting to estimate the parameters of the probe signal intensity distribution. The histogram of the logarithm of probe signal intensities has a nonsymmetrical shape and can be fitted either by a nonsymmetrical probability density function (pdf) or by a combination of symmetrical and/or nonsymmetrical pdfs. The histograms analyzed in this paper were fitted by the weighted sum of two log-extreme largest value pdfs. Increasing the number of log-extreme largest value pdfs (as well as combination of other pdfs) did not improve the data fitting significantly.

The second algorithm is nonparametric. It aligns the signal intensity distributions of the two arrays by using series of intervals of normalization. All signal intensities within the interval are normalized using linear transformation. This algorithm depends only on maximum, minimum values of arrays and the length of the interval of normalization. It should be noted that there are two extreme cases for this algorithm, where: (1) the interval of normalization is equal to the difference between maximum and minimum values; and (2) the interval of normalization is small and it contains $\leqslant 1$ values. The first case corresponds to the well-known method of "global" normalization when a single normalization factor is used for the whole set of microarray data. The linear relation between array intensities usually does not hold in general and the value of error Q after normalization will be not optimal. Correlation coefficient R for this extreme case will be exactly 1 for all genes. The second case is equivalent to the substitution of the values of the signal intensities for the array to be normalized by those for the model array with respect to the positions of values obtained after ordering. It provides the optimal values of Q for this method ($Q = 0$) but the average correlation coefficient R for genes is not optimal. It means that the optimal size for the interval of normalization exists and can be found using a weighted sum of Q and R as the objective function.

References

[1] Affymetrix Microarray Suite User Guide. Version 4.0 (2000).

[2] D.J. Lockhart, E.A. Winzeler, Genomic: gene expression and DNA arrays, Nature 405 (2000) 827–836.

[3] R.A. Young, Biomedical discovery with DNA arrays, Cell 102 (2000) 9–15.

[4] E.E. Schadt, C. Li, C. Su, W.H. Wong, Analyzing high-density oligonucleotide gene expression array data, J. Cell. Biochem. 80 (2000) 192–202.

[5] E.E. Schadt, C. Li, B. Ellis, W.H. Wong, Feature extraction and normalization algorithm for high-density oligonucleotiden gene expression array data, UCLA (2001).

[6] N.A. Hastings, J.B. Peacock, Statistical Distributions, Butterworth & Co. Ltd, London, 1975.